Analysis of Identifier Performance using a Deterministic Linkage Algorithm

Shaun J. Grannis MD, J. Marc Overhage MD PhD, Clement J. McDonald MD

Regenstrief Institute for Health Care, Indiana University, Indianapolis IN

ABSTRACT

As part of developing a record linkage algorithm using de-identified patient data, we analyzed the performance of several demographic variables for making linkages between patient registry records from two hospital registries and the Social Security Death Master File. We analyzed samples from each registry totaling 6,000 record-pairs to establish a linkage gold-standard. Using Social Security Number as the exclusive linkage variable resulted in substantial linkage error rates of 4.7% and 9.2%. The best single variable combination for finding links was Social Security Number, phonetically compressed first name, birth month, and gender. This found 87% and 88% of the links without any false links. We achieved sensitivities of 90% to 92% while maintaining 100% specificity using combinations of social security number, gender, name, and birth date fields. This represents an accurate method for linking patient records to death data and is the basis for a more generalized de-identified linkage algorithm.

INTRODUCTION

Because the information needed to answer important health research, management, and policy questions is usually scattered across many independent databases, methods for accurate linkage of patient records from independent sources are critical. Researchers have successfully used a variety of linkage methodologies[1,2].

Automated linkage methodologies are conceptually divided into two broad categories: deterministic and probabilistic.[3] Deterministic algorithms employ a set of rules based on exact agreement/disagreement results between corresponding fields in potential record pairs. Such algorithms are designed to match on a reliable identifier with high discriminating power and then perform verification using additional parameters. For example, linkage may be attempted using Social Security Number (SSN), which is then verified by first and last names.[1] If linkage is unsuccessful, one uses another composite key such as first and last name verified by other identifiers.

Probabilistic algorithms use statistical methods [2,4,5]. Frequency of identifier agreement and disagreement is derived from potential linked and non-linked record-pairs in the data sets. From this information, likelihood scores are calculated for each potential record-pair[5]. The likelihood scores for all potential record-pairs ideally form a bimodal distribution where low scores represent non-links, high scores represent probable links, and intermediate scores represent indeterminate links.

In addition to exact matching, methods exist for establishing agreement between fields such as approximate string comparison[6], phonetic encoding, and nearness metrics[7].

Although probabilistic methods may discriminate better than deterministic methods, in some cases their results require human intervention, and agreement likelihood information may not be readily available for all data.[8] Additionally, deterministic approaches often require less development time and still achieve acceptable results[1,3,4].

While much information can be gained from linked databases, steps must be taken to assure confidentiality of patient records.[9] We are developing a linkage method using data de-identified by a one-way hash function [10,17]. Nearness metrics cannot be used for data de-identified in this way because nearness information is lost in hash functions. Therefore, we must find other mechanisms to reduce variation that might otherwise be accounted for by nearness measures. It is important to avoid mechanisms that require human supervision, because that would break confidentiality in many circumstances, and the cost of supervised matching can be high. Consequently, we have implemented a deterministic, or exact match linkage method.

METHODS

This work was performed as part of the Shared Pathology Information Network (SPIN) project for which we received IRB approval. Using records from two hospital systems' patient registries, our goal was to maximize the chance for an individual to link to the Social Security Death Master File (SSDMF) even after applying a one-way hash function to all identifiers. This problem has general relevance to all medical databases and registries because a match to the SSDMF provides the best indicator of vital status (i.e. whether the patient is living or deceased). Mortality is an important outcome variable for many research questions[11] and we believe the SSDMF is the best source for that data.

The SSDMF is a publicly available database containing demographic data for over 65 million deceased individuals. A one-time snapshot can be purchased for $1,750 and monthly updates are available for $6,900 per year. The database has fields for SSN, name, date of birth, date of death, state or country of residence, ZIP code of last residence, and ZIP code of lump-sum payment. The Social Security Administration (SSA) receives approximately 90% of its death notifications from funeral homes, friends, and relatives of the deceased; postal authorities and financial institutions contribute another 5%. The remaining 5% are derived from computer matches with Federal and State agency data. The file is updated with additions, deletions, and modifications on a weekly basis.[12] The SSA maintains

that absence from the database is not proof the patient is alive because some deaths are not recorded. The CDC lists 2,391,043 decedents for 1999 compared to 2,154,018 (90.1%) included in the SSDMF for that year.

For this study we used patient registries from two hospitals in central Indiana. Hospital A is a public inner-city hospital system with a large Medicare/Medicaid population. Hospital B is a private urban hospital system that invested in extensive patient registry clean-up in 1999.

Selecting Indiana Death Records: Patient registries were obtained in December, 2001. We developed an Indiana subset of the SSDMF to speed up the matching process described below. An SSDMF record was included in this subset if any of the fields indicated the patient worked in, lived in, or obtained their SSN from Indiana using following data in the SSDMF: first 3 digits of SSN in the range 303-317; ZIP code for last residence or lump-sum payment ZIP code in the range 46000-47999; or an Indiana state of residence.

Preprocessing: Names and other variables can include variations and errors such that exact string matches may fail when a human reader might recognize them or the equivalent (e.g. "Jim" and "James"). To achieve de-identified matching, we plan to apply a one-way hash function to all fields before attempting linkage, and all information that could help in close matches will be lost. We thought that pre-processing names using a phonetic compression algorithm would help overcome such variations and errors. There are several phonetic compression algorithms; examples include Soundex[13], Metaphone, and the New York State Identification and Intelligence System algorithm (NYSIIS).[14] The NYSIIS algorithm has high discriminating power.[15] NYSIIS codes for first and last names were generated for each data set.

To eliminate last name, first name order reversal errors, we converted names from base 27 (A-Z) to base 10, summed them together, and re-converted to base 27. In this way "JOHN SMITH" and "SMITH JOHN" both produce the sum "SWYAV". We applied this same process to the NYSIIS-transformed first and last names.

Gender was available in the patient registries, but the SSDMF contains no fields for gender. When gender was missing from the hospital registration we imputed it using the non-intersecting names from the top 1000 male and female first names derived from 1990 U.S. Census data. We did the same for all SSDMF records.

Birth date and SSN are also subject to errors, but there is nothing analogous to Soundex-like rules for these variables. To accommodate errors in birth date, we decomposed it into month, day, and year variables; we used various combinations to attempt linkage. When SSN was erroneous we used other linkage criteria such as full name, birth date, and gender.

We preprocessed the data from each of the candidate match fields shown in Table 1. Because identifiers such as race, mother's maiden name, and institutional identifiers that were present in the hospital records were not present in the SSDMF, they were not included in matching rules. We used only the preprocessed variables in our analysis. In the context of anonymous linkage, we could perform this preprocessing at each source system before applying a one-way hash without compromising confidentiality. However, we examined the performance of both the raw and NYSIIS names. The preprocessing was intended to increase the chance of a correct match.

Manual Analysis: We developed a gold standard for measuring the error rates of the linkage variables and for comparing the matching accuracy of various combinations of these variables as follows. Using SSN as the single identifier, we linked the patient registries to the Indiana subset of the SSDMF resulting in potentially linked record pairs. If a hospital record linked to more than one record in the SSDMF, the first record pair was used. As the first stage, we obtained a random sample of n=1000 record-pairs from each institutions' potential links. The two samples were then manually reviewed and record pairs were labeled as correct or incorrect links.

Retrospective analysis of both 1000 patient samples revealed that all incorrect links based on SSN alone mismatched either on first names or birth years. In hospital A, the 84/1000 manually-labeled incorrect links were found among record pairs mismatched either on first name or birth year. Similarly, in hospital B, the 39/1000 incorrect links were found among record pairs meeting the same mismatch criteria.

To create a larger set of test cases, we took a random sample of 5000 record pairs linked by SSN alone from each hospital and manually reviewed all cases that mismatched on first name or birth date. Of the 5000 record pairs in each sample, 1,367 record-pairs from hospital A (27.3%) and 825 record pairs from hospital B (16.5%) were manually reviewed and labeled as correct or incorrect links. The n=1000 and n=5000 samples from each hospital were then combined to form gold standards of n=6000 record-pairs. We determined sensitivities and specificities for multiple combinations of candidate

Table 1: Preprocessed identifiers

Identifier	Values	Preprocessing Rules
Social Security Number (SSN)	0-9	Remove non-numeric characters; nullify if not 9 digits; nullify if not valid
Last Name (LN)	A-Z	Remove non-alphabetic characters, suffix and prefix nullify invalid names.
First Name (FN)	A-Z	Remove non-alphabetic characters, suffix and prefix nullify invalid names.
Name Sum (NS)	A-Z, zero	Produced after pre-processing of Last and First Names.
Gender (G)	M, F	If null, or ≠ (M F), attempt imputation from first name list based on census list.
NYSIIS encoding of Last Name (LNY)	A-Z, zero	Produced after pre-processing of Last and First Names.
NYSIIS encoding of First Name (FNY)	A-Z, zero	Produced after pre-processing of Last and First Names.
Sum of NYSIIS Names (SNY)	A-Z, zero	Sum of LNY and FNY
Month of birth (MB)	0-9	Convert from alphabetic month, 0 if < 0 or > 13
Day of Birth (DB)	0-9	0 if (< 0 or > 31)
Year of Birth (YB)	0-9	0 if (< 1800 or > 2001)

linkage variables within these gold-standard record pairs.

Non-SSN Linkage: For SSN record pairs labeled as incorrect links, we attempted a second linkage to the Indiana SSDMF using first name, last name, gender, and birth date. These were manually reviewed and labeled as correct or incorrect links. The correct links not generated by SSN were then compared to the initial incorrect SSN-generated links.

RESULTS

A substantial number of patient registration records, approximately 35%, lacked SSNs at each institution. Only the hospital records with valid SSNs were used in this study. When we linked these hospital records to the Indiana subset of the SSDMF, 57,446 (8.4%) of hospital A's records linked to a record in the Indiana SSDMF, and 147,878 (10%) records from hospital B linked.

We used the patient registry records that linked by SSN alone to the SSDMF to obtain the gold standard data set of 6000 record pairs. Among the 6000 gold standard record pairs, using SSN as the exclusive match variable, hospital A had 550 incorrect links, indicating a 9.2% SSN error rate, and hospital B had 281 incorrect links, indicating a 4.7% SSN error rate.

Table 2 shows the individual identifier mismatch rates among correct links based on SSN alone. Assuming that the SSDMF carries the correct information, these data provide an estimate of the error rates in the recorded information for each of the listed patient identifier fields. However, we cannot consider mismatches on first and last names to be strict errors because interchange between first names, nicknames and varying uses of first and middle initials confound this comparison. Further, the gender figures are not precise because all of the gender values in the SSDMF file are imputed.

Table 2: Identifier Error Rates Among Correct SSN-based Links

	Error Rates (%)	
	Hospital A (n=5450)	Hospital B (n=5719)
Last Name	5.9	2.1
First Name	12.5	8.2
Name Sum	16.7	9.9
NYSIIS Last Name	3.9	1.5
NYSIIS First Name	9.5	7.2
NYSIIS Sum	12.3	8.3
Gender	0.6	0.6
Month of Birth	3.7	1.8
Day of Birth	8.4	5.3
Year of Birth	8.2	4.2

There are some interesting observations we can make from this Table. Error rates were higher at hospital A as compared to hospital B, which had invested in a major clean up of their registration systems 3 years ago. It is notable that the month of birth is more accurate than year or day of birth. Also as expected, the NYSIIS algorithm had a lower mismatch rate than raw names. However the mismatch rate with NYSIIS was not zero, reminding us that phonetic transforms do not equivalence minor name differences like "Bill" and "Gill".

Among the record pairs not linkable by SSN, the use of name and birth date criteria identified an additional 196 correct links between hospital A and the Indiana SSDMF, while the same process identified another 109 correct links in hospital B. Using these links we analyzed the original SSN-linked record pairs for errors.

SSN errors consisted of three types shown in Figure 1. The most common error appeared to be due to spousal mix-ups (56% hospital A, 39% hospital B) in that a female of one record was linked to a male record sharing the same last name. Typographical errors (41% hospital A, 30% hospital B) and SSN collisions of unknown etiology (3% hospital A, 31% hospital B) accounted for the remainder of the errors. Figure 1 shows examples using fictitious data.

Figure 1: SSN Error Examples

	SSN	LN	FN	G	MB	DB	YB
Typographical Errors							
Institution Data	123456789	SMITH	FRED	M	12	20	1908
SSDMF Correct Name Link	123456789	SMITH	FRED	M	12	20	1908
SSDMF Incorrect SSN Link	123456789	JONES	PAT		7	13	1914
Spousal Linkage							
Institution Data	987654321	COLLINS	MARY	F	8	29	1917
SSDMF Correct Name Link	987708345	COLLINS	MARY	F	8	29	1917
SSDMF Incorrect SSN Link	987654321	COLLINS	TIM	M	4	3	1915
Unexplained Collisions							
Institution Data	586192376	PRATT	DAVID	M	10	3	1932
SSDMF Correct Name Link	666248764	PRATT	DAVID	M	10	3	1932
SSDMF Incorrect SSN Link	586192376	WILSON	ERIN		5	17	1940

The rows in Tables 3 and 4 describe sets of identifiers that could be used for linking patients and their corresponding false positive and false negative link rates. The best single combination of identifiers for finding matches was SSN, first name transformed by the NYSIIS, month of birth, and gender. This combination found 87% to 88% of the possible links without finding any false links. Taking the union of more than one set of keys – that is link by one set of keys, then link by another set of keys, and include all of the links from any of these steps in the final result – yielded an 89% to 90% link rate without picking up any false links. Adding links on first name, last name, and full birth date increased these yields to 90-92%.

DISCUSSION

Hospital registries contain substantial numbers of errors in SSNs that prohibit the use of SSN as a single linkage key. Additional fields have to be added to avoid incorrect links. Similar error rates in the SSN have been reported previously.[16] Nearly half of the SSN errors are due to spousal mix-ups, almost certainly due to a mix up between the guarantor's SSN and that of the patient, or beneficiary. Additional linkage identifiers such as gender and first name help to avoid incorrect links between beneficiaries and guarantors. We recommend that health care systems develop registration procedures to avoid the incorrect assignment of guarantor's SSN to a beneficiary.

Linkage criteria that include SSN combined with variables from both name and birth date maximize the match rate while keeping the false positive rate near zero. Identifier variations are not independent; people with the same last names may end up using the same SSN because of beneficiary or other errors. The first name and

Table 3: Results of 6,000 random samples taken from 57,446 record-pairs linked by SSN between Hospital A and SSDMF Indiana

	Links		Non-links			
Linked Identifiers	Correct	Incorrect	Correct	Incorrect	Sensitivity (%)	Specificity (%)
SSN Alone	5450	550	0	0	100	--
Name Criteria:						
SSN, LN, FN	4541	7	543	916	83.2	98.7
SSN, LNY, FNY	4775	7	543	675	87.6	98.7
SSN, SNY	4782	7	543	668	87.7	98.7
Date Criteria:						
SSN, MB, DB, YB	4557	2	548	893	83.6	99.6
Name/Date Criteria with SSN:						
SSN, FN, YB, G	4350	0	550	1100	79.8	100
SSN, FNY, YB, G	4496	0	550	954	82.5	100
SSN, FNY, MB, G	4724	0	550	726	86.7	100
Name /Date Criteria without SSN:						
LN, FN, MB, DB, YB, G	3996*	0	550	1650	70.1	100
Union of (FNY,YB,G), (FNY, MB, G), and (LN,FN,MB,DB,YB)	5053	0	550	593	89.5	100

* Potential links for non-SSN matches = 6196

Table 4: Results of 6,000 random samples taken from 147,848 record-pairs linked by SSN between Hospital B and SSDMF Indiana

	Links		Non-Links			
Linked Identifiers	Correct	Incorrect	Correct	Incorrect	Sensitivity (%)	Specificity (%)
SSN Alone	5719	281	0	0	100.0	--
Name Criteria:						
SSN, LN, FN	5157	2	279	562	90.2	99.3
SSN, LNY, FNY	5247	2	279	474	91.7	99.3
SSN, SNY	5245	2	279	474	91.7	98.9
Date Criteria:						
SSN, MB, DB, YB	5216	2	279	503	91.2	99.3
Name and Date Criteria:						
SSN, FN, YB, G	4997	0	281	722	87.4	100
SSN, FNY, YB, G	5048	0	281	671	88.3	100
SSN, FNY, MB, G	5181	0	281	538	90.6	100
Name and Date Criteria without SSN:						
LN, FN, MB, DB, YB, G	4776*	0	281	1052	81.9	100
Union of (FNY,YB,G), (FNY, MB, G), and (LN,FN,MB,DB,YB)	5331	0	281	497	91.5	100

* Potential links for non-SSN matches = 6109

gender provide important protections against such errors. Gender is included to avoid the theoretical possibility of an incorrect NYSIIS linkage between family members with similar first names who share SSN and birth date parameters.

The preprocessed linkage variables that perform reasonably well in this study are suitable for a de-identified linking mechanism. After being preprocessed at the local information system, identifiers can be encrypted via a secure one-way hash, using a one-time seed shared by all sites. The hashed keys can be sent to a trusted third party for linking and that party can assign random codes to each patient.[17]

We restricted the matching to the Indiana subset of the SSDMF to limit file size and computer time. To find all possible deaths in a local population of patients, one would link to the entire SSDMF. We would expect to find more links between patients in the registration files but also to encounter higher error rates, because the larger number of individuals in the target file would provide greater chances for links between different individuals who happen to have the same identifiers.

These results are based on modest sample sizes, and further analysis of larger populations is warranted. Our methods apply to decedent matches and patients from the Midwest. This may not generalize to other populations with high percentages of Hispanic or Asian names. By its nature, the death index contains an older population; linkage performance in a younger, more diverse population may differ. Further, assuming that the SSDMF file contains much cleaner data than the average hospital registration file, we would expect a lower link rate and more errors when data from both files are derived from patient registries.

This is an accurate method of linking patient records to death data, and will be the basis for a more generalized de-identified linkage algorithm. Future work includes linking registry data to the entire SSDMF to study the error properties and match rates using a larger data set. Work will also be directed toward improving non-SSN name matches. We will also consider use of some statistical properties such as name and birth date frequencies to improve matching precision.

REFERENCES

1. Potosky A, Riley G, Lubitz J, et al. Potential for Cancer Related Health Services Research Using a

Linked Medicare-Tumor Registry Database. Medical Care 1993;31(8):732-748.

2. Whalen D, Pepitone A, Graver L, Busch JD. Linking Client Records from Substance Abuse, Mental Health and Medicaid State Agencies. SAMHSA Publication No. SMA-01-3500. Rockville, MD: Center for Substance Abuse Treatment and Center for Mental Health Services, Substance Abuse and Mental Health Services Administration, July 2000.
3. Liu S, Wen SW. Development of Record Linkage of Hospital Discharge Data for the Study of Neonatal Readmission. Chronic Diseases in Canada 1999; 20(2):77-81.
4. Gill, L., Methods for Automatic Record Matching and Linking and their use in National Statistics. Her Majesty's Stationary Office, Norwich, 2001.
5. Fellegi, I.P., & Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), 1183-1210.
6. Porter E, Winkler W. Approximate String Comparison and its Effect on an Advanced Record Linkage System. Record Linkage Techniques--1997: Proceedings of an International Workshop and Exposition. National Academy Press, Washington DC 1999.
7. Sideli R, Friedman C. Validating Patient Names in an Integrated Clinical Information System. Symposium on Computer Applications in Medical Care, Washington, DC. November 1991:588-592.
8. Van Den Brandt PA, Schouten LJ, Goldbohm RA, Dorant E, Hunan PMH. Development of a record linkage protocol for use in the Dutch Cancer Registry for epidemiological research. Int J Epidemiol 1990; 19:553-8.
9. Department of Health and Human Services, Office of the Secretary. The Health Insurance Portability and Accountability Act of 1996, Standards for Privacy of Individually Identifiable Health Information; Final Rule. Federal Register 65 FR 82462; December 28, 2000. Available at: http://www.hcfa.gov/hipaa/hipaahm.htm
10. Burrows, JH. Secure Hash Standard. Federal Information Processing Standards, Publication FIPS PUB 180-1 <http://www.itl.nist.gov/fipspubs/fip180-1.htm> website accessed 3/1/2002.
11. Pates R, Scully W, et al. Adding Value to Clinical Data by Linkage to a Public Death Registry. MedInfo 2001;10(Pt 2):1384-8.
12. Social Security Administration, Office of the Inspector General, Unresolved Death Alerts Over 120 Days Old (A-09-00-10001). Audit Report; 2001 August.
13. Knuth DE. The Art of Computer Programming, Volume 3/Sorting and Searching, Second Edition. Addison-Wesley Publishing Company, 1998.
14. Lynch BT, Arends WL. Selection of a surname coding procedure for the SRS record linkage system. Washington, DC: US Department of Agriculture, Sample Survey Research Branch, Research Division, 1977.
15. Newcombe HE. Handbook of Record Linkage, Methods for Health and Statistical Studies, Administration, and Business. Oxford University Press, 1988.
16. Newman T, Brown A. Use of Commercial Record Linkage Software and Vital Statistics to Identify Patient Deaths. J Am Med Inform Assoc. 1997 May-June; 4 (3): 233-237.
17. Schadow G, McDonald CJ Maintaining Patient Privacy in a Large Scale Multi-Institutional Clinical Case Research Network. AMIA Proceedings (2002 Submission).

ACKNOWLEDGEMENTS

This work was performed at the Regenstrief Institute for Health Care in Indianapolis, Indiana and was supported in part by grants from the National Library of Medicine (T15 LM-7117-05), the National Cancer Institute (1 U01 CA91343-01), and The Indiana Genomics Initiative (INGEN) of Indiana University, which is supported in part by Lilly Endowment Inc.